Chunk Parsing and Entity Relation Extracting to Chinese Text by Using Conditional Random Fields Model

Junhua Wu, Longxia Liu

College of Electronics and Information Engineering, Nanjing University of Technology, Nanjing, China.
Email: wujh@njut.edu.cn

Received March 19th, 2010; revised July 20th, 2010; accepted July 30th, 2010.

ABSTRACT

Currently, large amounts of information exist in Web sites and various digital media. Most of them are in natural language. They are easy to be browsed, but difficult to be understood by computer. Chunk parsing and entity relation extracting is important work to understanding information semantic in natural language processing. Chunk analysis is a shallow parsing method, and entity relation extraction is used in establishing relationship between entities. Because full syntax parsing is complexity in Chinese text understanding, many researchers is more interesting in chunk analysis and relation extraction. Conditional random fields (CRFs) model is the valid probabilistic model to segment and label sequence data. This paper models chunk and entity relation problems in Chinese text. By transforming them into label solution we can use CRFs to realize the chunk analysis and entities relation extraction.

Keywords: Information Extraction, Chunk Parsing, Entity Relation Extraction

1. Introduction

At present, information is presented in various digital media. Many of them are organized in natural language, such as information in Web pages, text document in digital library etc. They are non structural or semi-structural and difficult to understand by computer. Further processing to the information is blocked. It makes large amounts of information wasted. So research on semantic Web, natural language understanding is developed in order to structure and retrieve information from Web pages or other natural language documents. And information extraction is important task in the work.

Information extraction is a process to retrieve information from large text set. It may be concerned with identifying named entity, extracting relationship and label properties of sentence etc. It is a subfield of natural language understanding. There are some methods for information extraction including methods based on rules [1,2] and statistical model [3-6].

Chunk analysis and relation extraction play the important roles in information extraction. It is a simplified syntax paring technology to define and label chunk based on syntax and semantics [7]. Comparing with full parsing this method only identifies the partial structure in a sentence, such as noun phrase or verb phrase. Through which, the simple syntax parsing can be implemented and information extraction may be more effective and simple.

The objective of entity relation extraction is identifying the relationship between entities in text. Miller et al. considered the problem of relation extraction in the context of natural language parsing and augmented syntactic parses with semantic relation-specific attributes [8]. It will be critical in events detecting and describing for research on information extraction. Entity relation may be explicit and implicit. Some encountered problems make studying on entities relation hard such as few dataset, difficult extraction of implicit relation and immature parsing to Chinese.

Conditional random fields model is a valid probabilistic model to segment and label sequence data [9]. In Chinese understanding, some research use CRFs in Chinese part-of-speech and word segmentation [10,11], but seldom in chunk parsing and entity relation extraction.

Compared with other statistical model CRFs can represent long-range dependences and multiple interacting features. Our innovation is that we analyze Chinese characteristics and then model chunk and entity relation problems as label problem. Moreover using CRFs real-
izes the chunk analysis and relation extraction.

2. Related Work

A number of approaches currently have been used for natural language tasks as part of speech tagging and entity extraction. They are usually based on rules or statistic models.

Text chunk divides a text in syntactically correlated parts of words. Steven introduced chunks [12] firstly. Many machine learning approaches, such as Memory-based Learning (MBL) [13], Transformation-based Learning (TBL) [14], and Hidden Markov Models (HMMs), have been applied to text chunking [15] for parsing.

Named entity is important linguistic unit. So there are many works such as named entity recognition, disambiguation, and relationship extraction on it [16-20]. The problem of relation extraction is starting to be addressed within the natural language processing and machine learning communities. Since it is proposed, many methods have been suggested. Methods based on knowledge base were used in decision of relation extraction firstly. But it is difficult to construct knowledge base. Therefore some methods based on machine learning were emerged, such as feature-based [16], kernel-based [17] method. Approach kernel-based is a valid one for relation extracting, but its training and testing time is long for large amounts of data.

2.1 Model for Information Extraction

A lot of research to information extraction is based on machine learning methods using statistic model because by the model sentence can be segmented and labeled. Statistical language model is a probability model which estimate probability of expected text sequence by computing probability. These models are concerned with Hidden Markov Model (HMM), Maximum Entropy Model (ME), Maximum Entropy Markov Model (MEMM) and conditional random fields model CRFs. Our method is also based on statistic model.

Hidden Markov models (HMMs) are a powerful probabilistic tool for modeling sequential data, and have been applied with success to many text-related tasks, such as part-of-speech tagging, text segmentation and information extraction [6]. HMM can be considered as a finite state machine that presents states and transition chains of an application. The model is built either by manual or training. Usually extracting text information is concerned with training and labeling. Maximum likelihood and Baum-Welch algorithm are used to learning sample data labeled or unlabeled. And then Viterbi algorithm is used to label state sequence with maximum probability in text needed processing.

HMM is easy to build. It needn’t large dictionary or rule sets with well flexibilities. There are many improved HMM model and their application in information extraction. Freitag and McCallum’s paper [3] uses stochastic optimization to search the fittest HMM. Souyama Ray and Mark Crave [4] choose HMM to represent sentence structure. Scheffer T, Decomain C and Wrobel S [5] proposes a method which uses active learning to minimize the label data for HMM training. But HMM is a generative model and independent hypothesis is needed, so it will ignore the context of information and lead to an unexpected result.

Maximum Entropy (ME) method [21] converts the sequence label into data classifying. Its principle can be stated as follows [22]:

1) Reformulate the different information sources as constraints to be satisfied by the target (combined) estimate.
2) Among all probability distributions that satisfy these constraints, choose the one that has the highest entropy.

The advantage of ME is [21]: It makes the least assumptions about the distribution being modeled other than those imposed by constraints and given by the prior information. The framework is completely general in that almost any consistent piece of probabilistic information can be formulated a constraint. Moreover, if the constraints are consistent, that is there exists a probability function which satisfies them, then amongst all probability functions which satisfy the constraints, there is a unique maximum entropy.

This ME method will lost sequence properties. So a model combining ME and MM (Markov Model) is emerged, that is MEMM [6].

In MEMM, the HMM transition and observation functions are replaced by a single function \( P(s'|s,o) \) that provides the probability of the current state \( s \) given the previous state \( s' \) and the current observation \( o \). In this model, as in most applications of HMMs, the observations are given—reflecting the fact that we don’t actually care about their probability, only the probability of the state sequence (and hence label sequence) they induce.

Conditional probability of transition between states is introduced in MEMM, which makes the arbitrary choice of properties possible. But MEMM is partial model which needs normalization for each node. Therefore only a localized optimization value is obtained. Also the problem named length bias and label bias [9] may be caused. It means the method will ignore those not in training dataset.

2.2 Label Bias

Classical discriminative Markov models, maximum entropy taggers (Ratnaparkhi, 1996), and MEMMs, as well as non-probabilistic sequence tagging and segmentation models with independently trained next-state classifiers are all potential victims of the label bias problem [9].
Consider a MEMM model shown in Figure 1 which is a finite-state acceptor for shallow parsing of two sentences:

The robot wheels Fred round.
The robot wheels are round.

Here [B-NP] etc. are labels for sentence. NP, VP, ADJP and PP mean Noun Phrase, Verb Phrase, Adjective Phrase and Prep Phrase. B or I stand for word location, begin or inter of a phrase.

It is obvious that sum of transition probability is 1 from a state 1 to other adjacent states. Because there is only one transition in state 3 and 7, while current state and observed value Fred are specified, conditional probability of next state is:

\[ p(4 \mid 3, Fred) = p(8 \mid 7, Fred) = 1 \]

But this equation will face to some problems if there isn’t an existing transition from state 7 to state 8 while observe value is Fred in training dataset. Generally a low probability is specified if an unknown event exists in training dataset. But for state with single output, the following equation have to be given:

\[ \sum_{\text{all states from state 7 to 8}} p(s \mid 7, Fred) = 1 \]

It means that the observed value Fred is ignored. This will result in that label sequence is not related to observed sequence. That is label bias.

Proper solutions require models that account for whole state sequences at once by letting some transitions “vote” more strongly than others depending on the corresponding observations [9].

Lafferty suggests a global model CRFs that can solve the problems discussed before. Instead of local normalizing CRFs can realize global processing, so a global optimization value will be produced. CRFs is a new graph model of probability which can represent the long-range dependences and multiple interacting features. Domain model of probability which can represent the long-range optimization value will be produced. CRFs is a new graph model that can compute conditional probability of output sequence when conditioned on input sequence of model.

Definition 1 [9]. Let \( G = (V,E) \) be a graph such that \( Y = (Y_v)_{v \in 2V} \), so that \( Y \) is indexed by the vertices of \( G \). Then \((X,Y)\) is a conditional random field in case, when conditioned on \( X \), the random variables \( Y_v \) obey the Markov property with respect to the graph: \( P(Y_v \mid X, Y_w, w \neq v) = P(Y_v \mid X, Y_w, w \sim v) \), where \( w \sim v \) means that \( w \) and \( v \) are neighbors in \( G \).

CRFs is a random field globally conditioned on the observation \( X \). if \( X = \{x_1, x_2, \ldots, x_n\} \) is specified as data sequence needed label then \( Y = \{y_1, y_2, \ldots, y_n\} \) is the result data which have been segmented or labeled by the model. The model computes the joint distribution over the label sequence \( Y \) given \( X \) instead of only defining next state in terms of current state.

The conditional probability of label sequence \( Y \) depends on the global interactional features with different weight.

Assume \( \Lambda = \{\lambda_1, \ldots, \lambda_k\} \) is a vector of features, conditional probability, for a given \( X \), \( P(\Lambda|X) \) is defined as follow:

\[
p_\Lambda(Y \mid X) = \frac{1}{Z_X} \exp \left\{ \sum_{i=1}^{T} \lambda_i f_i (y_{i-1}, y_i, X, t) \right\} \quad (1)
\]

\[
Z_X = \sum_{Y} \exp \left\{ \sum_{i=1}^{T} \lambda_i f_i (y_{i-1}, y_i, X, t) \right\} \quad (2)
\]

\( Z_i \) is a normalized value that makes the total probability of all state sequence is 1 for given \( X \). \( f_i (y_{i-1}, y_i, X, t) \) is a feature function to mark the feature at position \( t \) and \( t-1 \) for observed \( X \). Its value is between 0 and 1. \( \Lambda = \{\lambda_1, \ldots, \lambda_k\} \) is corresponding to the context of data sequence and is a weight set of \( f_i (y_{i-1}, y_i, X, t) \).

If we want to use the CRFs model to obtain expected result the critical task is training model. A model trained can produce optimization \( P(Y \mid X) \), that is \( Y^* = \arg \max_Y p(Y \mid X) \). It also means \( \Lambda = \{\lambda_1, \ldots, \lambda_k\} \) will be deter-

Figure 1. Finite-state acceptor for shallow parsing of two sentences

Copyright © 2010 SciRes.
Chunk Parsing and Entity Relation Extracting to Chinese Text by Using Conditional Random Fields Model

4. Chunk Analysis Based on CRFs

Chunk is firstly proposed by Abney [12]. He thinks chunk is the syntax element between word and sentence and with non-recursive properties.

Chunk analysis is partial parsing, also named shallow parsing, relative to complete parsing with simplified policy [15]. It is a new technology of natural language processing. Full parsing can produce a complete parse tree finally by series analysis process to sentence, which needs large cost. But chunk analysis only needs to identify some structures of the sentences such as non-recursive noun phrase, verb phrases etc, called chunk. By dividing sentence into different chunks in syntax or semantics and labeling chunks we can improve the efficiency of information extraction. It is a policy between lexical analysis and syntax analysis. Chunk partitioning and identifying are completed by chunk parsing in natural language processing.

4.1 The Definition and Label of Chunks

Definition 2. Chunk is a structure that is non-recursive phrase meet syntax. Each chunk has a head word and begins or ends at this word.

Non-recursive phrase means nested structure not exist. That is, all chunks are the same level.

In this paper chunk analysis and entity relation extraction will be converted into the label solution. Data sequence $X$ is made up of some words. For each word $W_0$ there are some words ahead or back of it. It is represented $W = \{W_0, W_1, W_2, \ldots, W_n\}$. $W_n$ stands for the last word of a chunk and $W_0$ is the nth one following $W_n$. $\Lambda = \{\lambda_1, \ldots, \lambda_k\}$, in model, can be thought as feature weights related to $W = \{W_0, W_1, W_2, W_3\}$. Each $\lambda$ is specified in model after training, and label sequence can be produced by Viterbi algorithm through running the model.

4.2 Model Training

CRFs model must be trained using labeled dataset to determine the model parameters. That trained model can be used to realize processing text which expects to be segmented and labeled. If $X$ is sentences that have been labeled and $Y$ is corresponding label sequence of chunk CRFs model training will make label sequence $Y^* = \arg \max_Y p(Y | X)$ optimal.

Here we use CRF++0.50 as training and testing tool. CRF++0.50 is a string learning tool based on CRFs principle. The training sample file and feature template file are needed in training process. Training will result in a CRFs model which will be used in labeling chunk to Chinese text.

<table>
<thead>
<tr>
<th>Chinese</th>
<th>English</th>
<th>IOB1</th>
<th>IOBES</th>
</tr>
</thead>
<tbody>
<tr>
<td>因而</td>
<td>So</td>
<td>I-CONJP</td>
<td>S-CONJP</td>
</tr>
<tr>
<td>我们</td>
<td>we</td>
<td>B-NP</td>
<td>S-NP</td>
</tr>
<tr>
<td>可能</td>
<td>may</td>
<td>B-VP</td>
<td>B-VP</td>
</tr>
<tr>
<td>会</td>
<td>be</td>
<td>I-VP</td>
<td>I-VP</td>
</tr>
<tr>
<td>面临</td>
<td>face</td>
<td>I-VP</td>
<td>E-VP</td>
</tr>
<tr>
<td>一个</td>
<td>a</td>
<td>B-MP</td>
<td>S-MP</td>
</tr>
<tr>
<td>不</td>
<td>un</td>
<td>B-NP</td>
<td>B-NP</td>
</tr>
<tr>
<td>稳定</td>
<td>stable</td>
<td>I-NP</td>
<td>I-NP</td>
</tr>
<tr>
<td>时期</td>
<td>period</td>
<td>I-NP</td>
<td>E-NP</td>
</tr>
</tbody>
</table>

Table 1. An example of label chunks
The training sample file is made up of some blocks and each block represents a sentence. The block form of training sample file is presented in Table 2.

There is a blank row between blocks. Each block includes some tokens and each token is a label word in one row. First column is the Chinese word and next column is the English word to help understanding. Third column lists the properties of the word (may be more than one column). Last column is the tag notations.

In section 3 we know some feature weights used in representing context of a word $W_0$. So it is important to select feature set. Generally context of a word and their properties are very useful for decision of feature. That means we can use some words which are previous or succeed to word $W_0$ as features. Features may be N-gram. Features $\{... W_2, W_1, W_0, W_0, W_1, W_2\}$ is named Uni-gram basic features and $\{... W_2, W_1, W_0, W_0, W_1, W_1, W_2\}$ is named Bi-gram basic features. Here $W_0$ stands for a word. In addition, advanced features $\{... WP_2, WP_1, WP_0, WP_1, WP_2\}$ which combine the word and its property together are also used to improve result of analysis. $P$ means property of a word. Table 3 shows various features of word “赤字” (deficit) in Table 2. The last column of Table 3 is English word corresponding to Chinese word.

So an observing window of token $W_0$ need to be given for training. The window includes $W_0$ and some words before and after it, that is $W = \{W_0, W_{(n-1)}, ..., W_0, ..., W_n, W_n\}$. Table 4 is an example about observing window. It is used as feature source for training vector $\Lambda = \{\lambda_1, ..., \lambda_k\}$.

Larger window provides more context feature, but it will increase the cost of processing. Too small window may lose important features. So we define windows size as 5, that is $W = \{W_2, W_1, W_0, W_1, W_2\}$.

The template file of features defines the feature item for training. After training $\Lambda = \{\lambda_1, ..., \lambda_k\}$ is produced, that is CRFs model has been available.

5. Entity Relation Extraction

Entity is the basic element in natural text, such as place, role, organization, thing etc. Entities play important roles in natural language text. Generally there are some relationships between them. Such as locating, belong to, adjacent and so on. These relationships may be explicit or implicit. Implicit relationship needs reasoning by knowledge. Entity relation extraction is the process of identifying the relationship between entities in text and labeling them. It is not only an important work in information extraction but also useful in automatic answer or semantic network.

Testing from MUC shows that many systems are able to process named entity to large of English document [26]. But entity relation extraction to Chinese may be difficult. As we known machine learning is the valid method for extracting, but it needs dataset labeled. Currently, “People’s Daily” labeled by Beijing university is perhaps a better choice. This dataset has been labeled in...
part-of speech, properties of word, named entity. But extracting implicit relationship needs more external knowledge.

5.1 The Definition and Label of Entity Relation

ACE 2006 defines six classes and 18 subclasses relationships between two entities. They are shown in **Table 5**. The dataset provided by ACE covers English, Chinese and Arabic.

This paper focuses on Chinese entity relation. Here we only illustrate two kinds of relationship of physical class, and only implement extraction based on the definition of these two relationships.

Definition 3: Relation 1 (M, N) is defined as located relationship. With Entity M, N ∈ geographical entity and N ⊆ M.

Definition 4: Relation 2(M, N) is defined as near relationship. With M, N ∈ geographical entity and 2(M, N) = 2(N, M).

In this paper the same principle as chunk analysis is used to realize extraction which training CRFs model through label sample dataset. Then using CRFs model realizes the extraction. Here we suggest nine kinds of notation to label the position relationship in dataset. **Table 6** presents these nine notations.

In the table each row presents one or two entities and their relationship. For a notation 1-B, 1 stands for Relation 1 (M, N) and B stands for the first entity M. The third column lists the sentence including entities. The end column shows the instance corresponding to notation.

### Table 5. Named entity relation types and instances

<table>
<thead>
<tr>
<th>Class</th>
<th>Subclass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical</td>
<td>Located, Near</td>
</tr>
<tr>
<td>Part-Whole</td>
<td>Artifact, Geographical, Subsidiary</td>
</tr>
<tr>
<td>Personal-Social</td>
<td>Business, Family, Lasting-Person</td>
</tr>
<tr>
<td>ORG-Affiliation</td>
<td>Employment, Founder, Ownership, Student-Alum, Sports-Affiliation, Investor-Shareholder, Membership</td>
</tr>
<tr>
<td>Agent-Artifact</td>
<td>User-Owner-Investor-Manufacturer</td>
</tr>
<tr>
<td>General-Affiliation</td>
<td>Citizen-Resident, Religion, Ethnicity, Org, Location</td>
</tr>
</tbody>
</table>

### Table 6. Nine kinds of labeling and instances

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
<th>Sentences for example</th>
<th>Instance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-B</td>
<td>Entity M in Relation1(M,N)</td>
<td>中国首都北京 (Beijing of China capital)</td>
<td>中国 (China)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>中国首都北京 (Beijing of China capital)</td>
<td>北京 (Beijing)</td>
</tr>
<tr>
<td>1-E</td>
<td>Entity N in Relation1(M,N)</td>
<td>城市北京和天津 (City Beijing and Tianjin)</td>
<td>北京, 天津 (Beijing, Tianjin)</td>
</tr>
<tr>
<td>2</td>
<td>Entity M,N in Relation 2(M,N)</td>
<td>位于中国首都北京的西单 (Xidan in Beijing of China capital)</td>
<td>北京 (Beijing)</td>
</tr>
<tr>
<td>1-E-1-B</td>
<td>Entity N in Relation 1(M,N), 1(N,S)</td>
<td>位于中国首都北京的西单 (Xidan in Beijing of China capital)</td>
<td>北京 (Beijing)</td>
</tr>
<tr>
<td>1-E-1-E</td>
<td>Entity S in Relation 1(M,N), 1(N,S)</td>
<td>位于中国首都北京的西单 (Xidan in Beijing of China capital)</td>
<td>西单 (Xidan)</td>
</tr>
<tr>
<td>1-B-2</td>
<td>Entity M, S in Relation 1(M,N), 2(M,S)</td>
<td>美国总统访问中国 (The president of America visits China)</td>
<td>美国, 中国 (America, China)</td>
</tr>
<tr>
<td>1-E-2</td>
<td>Entity N,S in 1(M,N), 2(N,S)</td>
<td>中国城市北京和天津 (City Beijing and Tianjin of China)</td>
<td>北京, 天津 (Beijing, Tianjin)</td>
</tr>
<tr>
<td>2-1-B</td>
<td>Entity N in 2(M,N), 1(N,S)</td>
<td>美国总统访问中国 (The president of America visits China)</td>
<td>中国 (China)</td>
</tr>
<tr>
<td>2-1-E</td>
<td>Entity S in 2(M,N), 1(N,S)</td>
<td>美国总统访问中国 (The president of America visits China)</td>
<td>北京 (Beijing)</td>
</tr>
</tbody>
</table>
Table 7. Experiment result of entity relation extraction

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CRFs trained</th>
<th>Time (s)</th>
<th>Precision (%)</th>
<th>Recall R(%)</th>
<th>F-Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>Model 1</td>
<td>48</td>
<td>39.8</td>
<td>45.3</td>
<td>42.4</td>
</tr>
<tr>
<td>M2</td>
<td>Model 2</td>
<td>112</td>
<td>38.2</td>
<td>57.2</td>
<td>45.8</td>
</tr>
<tr>
<td>M3</td>
<td>Model 3</td>
<td>203</td>
<td>45.3</td>
<td>59.6</td>
<td>51.5</td>
</tr>
<tr>
<td>M4</td>
<td>Model 4</td>
<td>293</td>
<td>50.2</td>
<td>65.5</td>
<td>56.8</td>
</tr>
<tr>
<td>M5</td>
<td>Model 5</td>
<td>547</td>
<td>63.1</td>
<td>69.8</td>
<td>66.3</td>
</tr>
<tr>
<td>M6</td>
<td>Model 6</td>
<td>923</td>
<td>73.1</td>
<td>74.9</td>
<td>74.0</td>
</tr>
<tr>
<td>M7</td>
<td>Model 7</td>
<td>1982</td>
<td>76.9</td>
<td>84.7</td>
<td>80.8</td>
</tr>
<tr>
<td>M8</td>
<td>Model 8</td>
<td>2439</td>
<td>87.9</td>
<td>89.4</td>
<td>88.6</td>
</tr>
<tr>
<td>M9</td>
<td>Model 9</td>
<td>3010</td>
<td>90.5</td>
<td>95.8</td>
<td>93.1</td>
</tr>
</tbody>
</table>

The data in table is the result of using M10 to test each model trained through M1-M9. Experiment shows that P, R and F increase with the more dataset used in training model. When we use M1 training the model F-Score is 42.4%. This is a disappointed value. But it only uses few sample dataset (newspaper of one day) for training. When using data of nine days we have obtained P,R,F values as 90.5%, 95.8% and 93.1% by use Model10. It illustrates it’s a valid method. If we provide enough sample dataset we may win better result.

6. Conclusions

This paper discusses the information extraction of Chinese text based on CRFs which aims at the chunk parsing and relation extraction. Processing Chinese text is a complex system. This is an exploration because we haven’t enough sample dataset for training CRFs model by now. But we think it’s an effective method by experiments since CRFs model possesses working with global features.

At present we are developing a prototype of information extraction so a lot of work will be continued. Absence of training database is the common problem for many kinds of language. But manual label will be large cost. Therefore there are some researches on automatic or semi-automatic constructing dataset. In addition, how to select suitable feature set and improve precision is also future works.

REFERENCES


